ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

Lung Cancer Detection using Transfer Learning and EfficientNet B2 Architecture

Ankita Tiwari

Department of Engineering Mathematics, College of Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India. tdrankita@gmail.com

Abstract

The world's most common cancer-related death is lung cancer, and better patient outcomes depend on early identification. The preferred imaging technique for screening for lung cancer is computed tomography (CT); however, due to its limited sensitivity, it can potentially miss diagnosis and produce false-negative results. Some current methods employ various filter types (Gabour, Median, etc.), which might not offer the necessary precision. So, to boost accuracy and processing speed, a transfer learning (TL) method is utilized with the help of EfficientNet B2. The deep learning-based strategy for tumor identification on CT scan pictures is proposed in this paper. The suggested technique automatically extracts features from CT images and categorizes them as normal or abnormal using Transfer learning a convolutional neural network (CNN).

Keywords—Lung cancer, CT scan images, Deep Learning, Data augmentation, Transfer Learning, CNN, EfficientNet B2.

I. INTRODUCTION

Lung cancer is the most significant cause of cancer deaths globally. Early identification is critical for treatment Lung cancer stands as a leading cause of global cancer-related fatalities, emphasizing the critical importance of early detection for successful treatment. Traditional approaches, often invasive and less successful, have been overshadowed by the advancements in machine learning and computer vision [1-2]. These innovations provide a non-intrusive, efficient, and precise path for diagnosing lung cancer. Transfer learning (TL), leveraging pre-trained models, enhances the performance of deep learning models in lung cancer diagnosis [3]. Application of TL to medical imaging, specifically using EfficientNet B2, capitalizes on its pre-training on a diverse image dataset, enabling it to learn general features beneficial for

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

various image recognition tasks. Adopting efficient Bayesian networks (BNets) further ensures reliable and interpretable outcomes with reduced computational complexity [4-5]. Ongoing research in the field calls for an extensive review of recent literature on deep learning and TL to support contemporary healthcare practices. By integrating these strategies, researchers have developed more accurate and efficient models for diagnosing lung cancer through medical imaging data [6-8]. These models, utilizing X-rays or CT scans, enable the early identification of lung cancer symptoms, leading to improved patient outcomes and reduced healthcare costs—an imperative area for study and development [9-10]. Consequently, the proposed research aims to scrutinize current techniques, approaches, and methods in deep learning and TL for health monitoring [11-14]. Deep learning, employing cascading layers of related functions, emerges as a powerful method for mining vast healthcare data to diagnose, treat, and prevent diseases and conditions [15-17]. While deep-learning applications may seem daunting to the layperson, industry experts acknowledge its global impact in addressing human challenges across diverse fields [18-21]. The evolution of lung cancer treatment commenced with the identification of therapeutically relevant genetic changes, notably the epidermal growth factor receptor (EGFR) mutation and the Anaplastic Lymphoma Kinase (ALK) fusion [22-25]. Initially perceived as mutually exclusive, it is now recognized that cancers may harbor multiple genetic drivers, such as coexisting EGFR and ALK mutations. A case series of Non-Small Cell Lung Cancer (NSCLC) patients with dual EGFR and ALK mutations is presented [4], [26-28]. The efficacy of immune checkpoint inhibition in fusion-driven NSCLC remains an unresolved issue, contributing to its status as a major cause of cancer-related death [29-31]. Recent evidence suggests the inefficacy of immunotherapy in fusion-driven NSCLC [5], [32-34]. Preprocessing and training medical data, including X-rays and CT scan images, involve various deep networks and machine learning techniques [35-39]. The subsequent sections of this work delve into existing methods in the literature, highlight key challenges, discuss the dataset and proposed methodology, present results, and provide a conclusion along with avenues for future research in Sections 3 through 5.

II. LITERATURE

Nowadays, most people are suffering from different diseases. In that, cancer plays a leading role. Cancer is a dangerous disease in which the body's cells multiply quickly and

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

uncontrollably, leading to abnormal growth. There are distinct types of cancers. The second most frequent cancer disease, conforming to reports, is lung cancer. To diagnose lung cancer by observing parameters like heart rate, blood pressure, etc., or using CT scan pictures. There are several working models for lung cancer detection. The major problems come with CT scan images processing in lung cancer detection. Onur et. al. proclaimed there are several ways to process these images. One such way is the OpenCV method. These images are converted into grayscale, then classification is done, but this method has deviations resulting in an undetectable result. After image processing, classification is another major issue that must be carefully done [6].

According to Senthil et. al., classification can be achieved through KNN (K Nearest Neighbors). KNN is a straightforward algorithm because all that needs to be computed is the distance between distinct points utilizing data for various attributes. However, it only works well with small datasets since calculating the distances between each data element would be too expensive. The main problem with KNN is the data clustering, and KNN's memory requirements, slow classification, and estimation are its drawbacks [7]. Besides Neal et. al., classification can also be done by CNN. CNN is a form of the deep network model used for tasks such as image segmentation and image data analysis. In issues involving picture recognition, CNN has very high accuracy. CNN does not retain object location and orientation. Inability to be spatially independent of the incoming data. A large volume of training information is necessary. The division of a picture into parts is known as image segmentation [8]. Through Fan et. al., Image segmentation is frequently used to identify boundaries and objects in photographs. Image segmentation generates noisier data. Because of the noise in the data, processing it takes longer. An image classification strategy based on the transform and the segmentation algorithm can be applied to address the issues. Overfitting occurs when your model matches the training dataset too closely [9].

Based upon the literature survey these are number of the key challenges that still needs to be addressed in future that are listed below:

- **Deviations in OpenCV**: The detection result of the OpenCV method has a significant variation. ANN is used in OpenCV for image classification. ANN's main issue is converting 2-D images to 1-D image vectors. This multiplies the number of training parameters. Improving trainable parameters necessitates additional storage and processing power [6].
- **Clustering of data**: K- means has difficulty grouping data, and groups fluctuate in size

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

and density; this happens when all the clusters are spherical, have equal radii, and are well separated. So, classification could be preferred over clustering [7].

- **Image segmentation**: Image segmentation creates more noisy data which further increases the processing time. An image edge detection based on a combination of the wavelets and the segmentation method can be used to address the issues [9].
- **Overfitting issues**: Dense Net has overfitting issues i.e.; they have inability to depict traits exhibited. Dense Net requires heavy GPU memory due to concatenation operations [10].
- **Pooling Loss Function**: In a Spatial Pyramid Pooling loss function cannot converge; it takes time to gather all images of the same size in a batch. It can be overcome by taking images of fixed dimensions [11].
- **High Dimensionality**: Gabor filter takes too much time to perform features extraction due to the high dimensionality of the feature vector. It can be prevented through CNN [12].
- **Lower throughput**: iTRAQ technique has a lower throughput and it is time-consuming. But it can be overcome using deep learning models[13].
- **Latent dynamic analysis**: Latent dynamic analysis (LDA) is not an efficient representation and works well only if the design is balanced. It cannot be entirely analyzed due to limited access to the network [14].
- **External fragmentation**: It is challenging to allocate contiguous memory to variable-sized partitions due to the Red Fox algorithm [15].
- **Training period**: The drawback of the ResNet algorithm is that a deeper network usually requires weeks for training. It takes more time to train because of vanishing and exploding gradient problems [16].

III. DATASET &METHODOLOGY

In this work, dataset contains CT scan images obtained through online resource [18]. In this dataset, 3 different sets of images are reserved, which are train, test, and valid. And further these images are classified into adenocarcinoma, large cell carcinoma, normal, and squamous cell carcinoma. This dataset contains an overall count of 1000, where 338 images of adenocarcinoma, 187 images of large cell carcinoma, 215 images of normal and 260 images of squamous cell carcinoma.

For quick array and data manipulation, NumPy and Pandas are essential data science libraries used in this work. And deep learning frameworks such as Keras and TensorFlow are

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

employed for the implementation. By using EfficientNet B2 from Keras, the image data generators make image data preprocessing easier. EfficientNet B2 is a variation of the EfficientNet model architecture designed specifically for image classification applications. It has been demonstrated that it achieves futuristic performance on a variety of picture classification benchmarks while remaining relatively tiny and computationally fast in comparison to other strong models.

During training, model checkpointing preserves the highest performing model. When model performance stops increasing, early halting ceases training. Thereafter, Transfer learning (TL) is introduced for fine-tuning the previously taught model. Efficient Net is a cutting-edge computer vision technique that is employed to perform scaling and depth- wise separable convolutions for high accuracy with fewer parameters. TL can lead to shorter training times and higher accuracy. The mathematical formulation for the TL is as also briefly discussed below:

A domain $\mathfrak{D} = \{X, P(X)\}$ is defined by two components: A feature space $X$and a marginal probability distribution $P(X$

) where $X = \{x_1, x_2, x_3, \ldots, x_n\} \in X$.

If two domains are different, then they either have different feature spaces ($X_t \neq X_s$) or different marginal distributions ($P(X_t) \neq P(X_s)$). Where is X_s is source domain and X_t is target domain.

Given a source domain $\mathfrak{D}s$ and corresponding learning task $\mathcal{T}s$, a target domain $\mathfrak{D}t$ and learning task $\mathcal{T}t$, **transfer learning** aims to improve the learning of the conditional probability distribution $P(Yt|Xt)$ in $\mathfrak{D}t$ with the information gained from $\mathfrak{D}s$ and $\mathcal{T}s$, where $\mathfrak{D}t \neq \mathfrak{D}s$ or $\mathcal{T}t \neq \mathcal{T}s$. After considering the above definition of domain and task, then we will have either $\mathfrak{D}t \neq \mathfrak{D}s$ or $\mathcal{T}t \neq \mathcal{T}s$, which results in four common transfer learning scenarios. The use of TL in the proposed work produces better results more quickly and the pre-training is also practiced. The proposed work overcame the above-mentioned image segmentation barrier by using data augmentation, which produced more precise data and corrected the overfitting issue. Since the problem has already been trained for a similar task, by starting with a pre-trained EfficientNet B2 model, transfer learning can be used to adapt the model to a new, related image classification task with less data and computation than would be required to train a model from scratch.TL provides a more significant learning rate during training. Improved

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

accuracy following training TL allows a ML model to merge at a higher performance level by providing a better point of reference and a higher learning rate, resulting in more accurate output [17].

Firstly, the input will be loaded and clustered into training and testing datasets. The clustered data will then be further divided into four distinct types of datasets. Each dataset is handled with average of 200 photos via data augmentation. EfficientNet in Keras is implemented for computation and training the model.

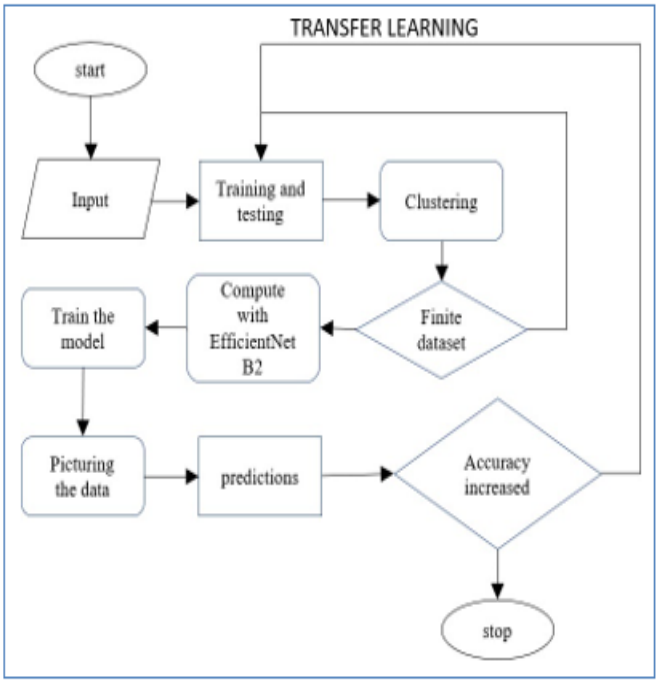

Fig:1. Proposed model

The training loss, validation loss, training accuracy, and validation accuracy graphs show the ratio of the functions calculated. The model then generates predictions with the necessary accuracy, and the resulting accuracy is handed back to the application for calculating with improved results using the TL which yields the maximum precision for the model. The proposed model data flow and operational process is shown in Fig:1. The model is developed using the transfer learning by CNN. The model will process the training data, and the testing dataset will be used for evaluation. The results are shown using graphs and a confusion matrix. The proposed model has the benefit of reusing the data again. The pseudo-code is as follows:

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

```
1. Import Needed Modules
2. Read input data and generate data of image paths and class
labels
3. IF the training set is finite
4.          Create training, testing, and validation generators
5.          Make a function to picture the Training Images
6.          Create the Model FOR 4 Types of Classes
7.                    Create a custom KERAS callback to continue
                                                  or stop and Compute data
                    with EfficientNet B2on 4 classes.
8.                    Instantiate custom callback
                      END FOR
9.          Train the model
10.         INITIALISE the function
                      display the training data.
11.         Make predictions on a test set
                      generate Confusion Matrix
                      Classification Report
12          Save the model.
   END IF
13          Clear the session and the graph
14          Repeat steps 2 through 12
```

IV. RESULT AND DISCUSSION

The input consists of a collection of CT scans that will be further classified into four classes: adenocarcinoma, normal, squamous cell, and large cell. Processing uses data augmentation to train and test the input.

- **Adenocarcinoma:** The most typical histologic form of lung cancer is adenocarcinoma. It is a malignant tumor classified as a non-small cell carcinoma of the lung that expresses glandular differentiation or mucin production in various patterns and levels of differentiation.
- **Squamous cell:** Non-small cell lung cancer sometimes referred to as squamous cell lung cancer, is characterized by squamous cell carcinoma (SCC) of the lung (NSCLC). Squamous cell lung tumors frequently develop in the lung's middle or a significant airway, like the left or right bronchus. Smoke from cigarettes is a critical factor in cellular change.

The sample input images for of adenocarcinoma and squamous cells of the lung cancer is shown in Fig:2.

- **Large cell:** Because of the size and appearance of its cells, this form of lung cancer is so termed. The lungs contain these cells all around. Compared to other types of non-small cell lung cancer, they also tend to develop

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

and spread more quickly. 10% to 15% of all non-small cell lung cancers are LCLCs.

- **Normal:** Similar to the TSS, each lobe may get a CT score between 0 and 5, with 0 representing no engagement, 1 representing less than 5% involvement,

2 representing 25% involvement, 3 representing 26-

49% involvement, 4 representing 50–75% involvement, and 5 representing involvement more than 75%. The overall CT score, which ranges from 0 to 25 points, was calculated by adding the values from each lobe.

The sample input images for of large and normal cells of the lung cancer is shown in Fig:3.

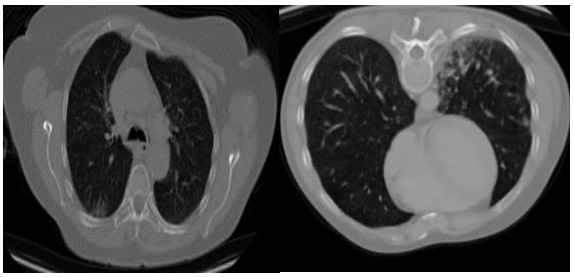

Fig:2. Sample images of adenocarcinoma (on the left) and squamous cells (on the right) of the lungs.

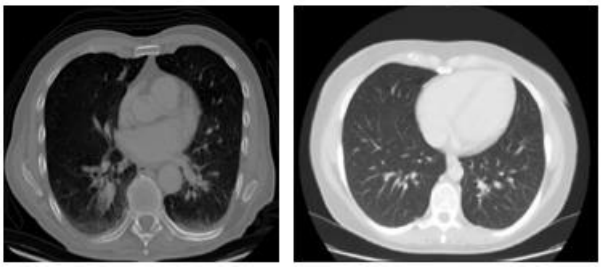

Fig:3. The above images represent Large cell carcinoma of the lungs on the left and a Normal CT Scan on the right.

In the proposed model, overfitting is prevented by adopting the early stopping strategy. The advantages of utilizing TL in the proposed model is to save time and resources. The proposed model is evaluated based on following parameters:

- **Precision**: Precision lets us visualize the machine learning model's dependability in identifying the model as positive.
- **Recall**: We can ignore negative samples and only calculate ones categorized as positive because it is independent of how the negative samples are categorized in the model.
- **F1-Score**: The F1 score gauges the accuracy of a classifier. It is the Precision and recalls harmonic mean. The classifier performs better at differentiating between positive and negative instances the higher the F1 score.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

- **Support:** It suggests that the evidence is found in the number of instances of each specific type that appear relevant such as responses in your test set. It can be calculated by adding the confusion matrix's rows.

These metrics concur with the aforementioned conclusion that this method is appropriate for accurately categorizing medical imaging. Theproposed model is developed without the use of ROI extraction (preprocessing) as it was producing poorer accuracy. In Fig:4, the dataset is clustered into three different datasets training, test, and validation datasets where each group has its own range. Adenocarcinoma has 195 train, 120 test, 23 validation images. Large cell has 115 train, 51 test, 21 validation images. Normal 148 train,54 test, 13 validation images. Squamous cell 155 train, 90 test, 15 validation images. The typical picture has a width of 434 and a height of 297. The model rounds each class up to a total of 200 photos via data augmentation. The validation and training loss values provide a better understanding of how the training efficiency changes with the number of epochs. Fig: 5 presents the loss function ratio of training and validation. They assist us in diagnosing any learning issues that may result in an underfit or an overfit model.

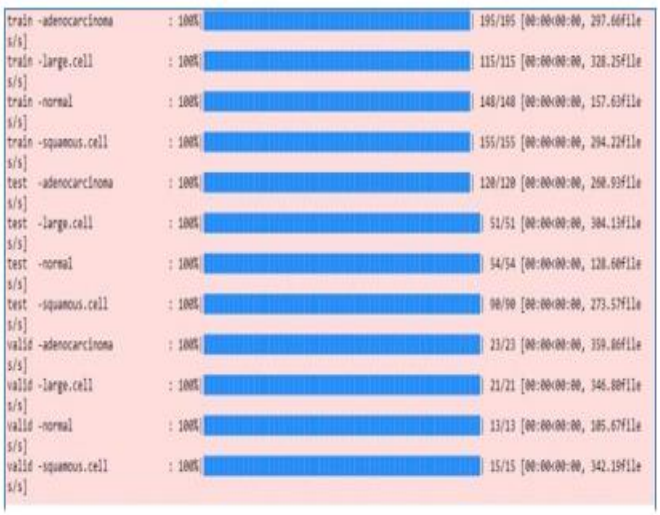

Fig:4. Train, Test and Validation data

In Fig:5, the green curve represents the validation loss, the red curve represents the training loss, and the best epoch is valued at 15.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper

 UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

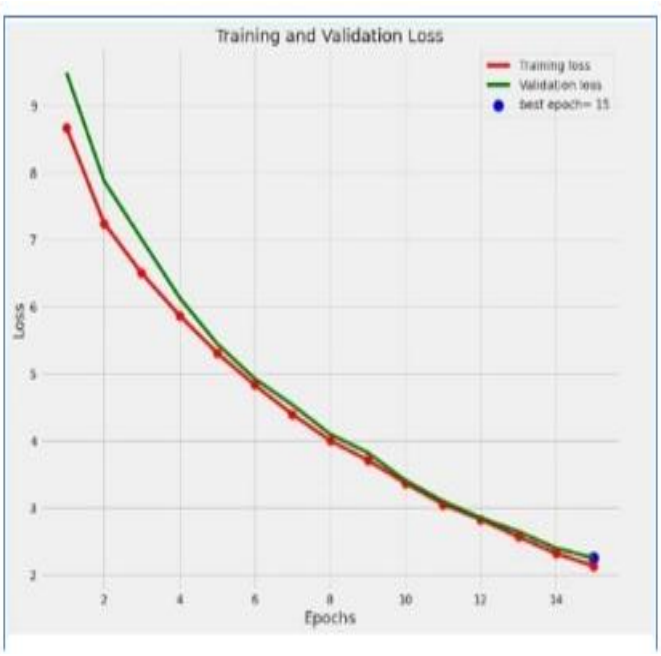

Fig:5. Loss function ratio of training and validation process.

Fig:6. Accuracy graph between training and validation.

Fig:6 depicts the accuracy graph between training and validation. How the model was trained on the training sample to categorize the images. How well the model can categorize the pictures using the validation dataset is known as valid accuracy. In Fig: 6. the green curve indicates the validation accuracy, and the red curve goes with the training accuracy. The finest epoch in the graph above may be found at 10. The following Fig: 7 shows the comparative analysis of the actual and predicted outcomes.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

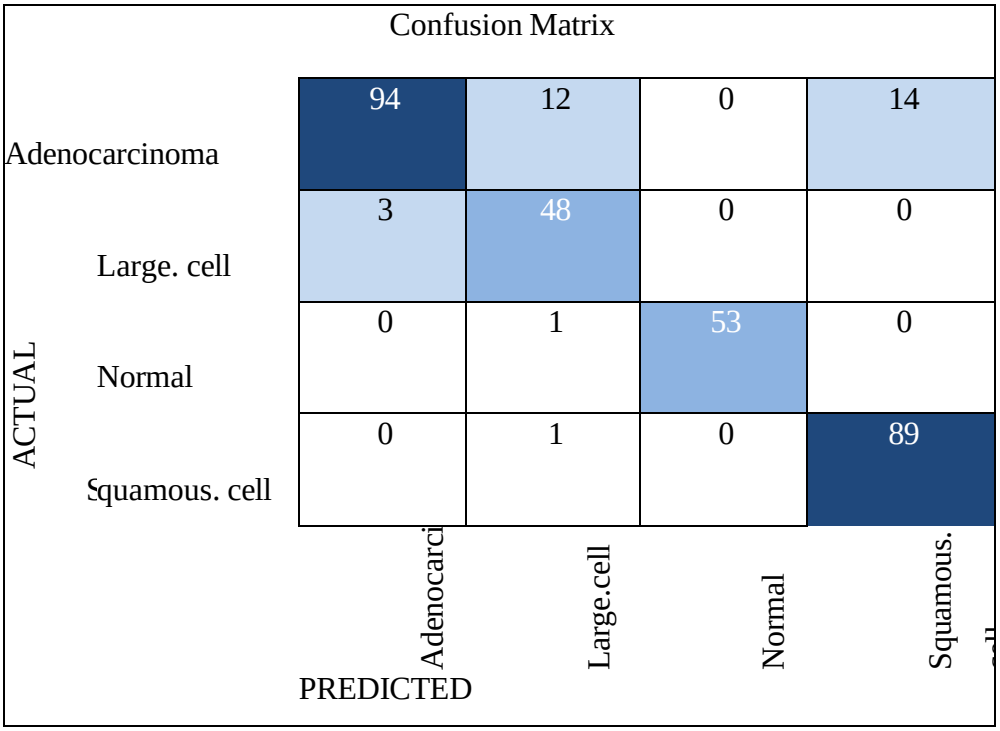

Fig:7. Confusion Matrix

Each row of the confusion matrix represents the examples in an actual class, while each column represents the instances in a predicted class, or vice versa. A confusion matrix is a special table arrangement that permits the display of the performance of an algorithm. The diagonal blocks with dark color in the above matrix stand in for the overall number of accurate predictions for each class.

Here, Precision is given by the ratio of true positives and the total no. of positives.

Precision = TruePositives / (TruePositives + FalsePositives)

i.e. precision of adenocarcinoma = 94/(94+3) =0.969.

A recall is given by the ratio of TruePositives and TruePositives + FalseNegatives.

Recall = TruePositives / (TruePositives + FalseNegatives). F1-Score is given by:

F1 Score = 2 * (Precision * Recall) / (Precision + Recall).

i.e. F1 score of adenocarcinoma = 2* (0.9696 * 0.7833) / (0.9696 + 0.7833) = 0.866.

Support is the sum of all TruePostive and FalsePositives.

i.e. support of adenocarcinoma = 94 + 12 + 0 + 14. Therefore, we have achieved a total accuracy of 90.16. The complete classification report is given below:

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

	Precision	*Recall*	*F1-Score*	*Support*
adenocarcin oma	0.9696	0.7833	0.8667	120
Largecell	0.7742	0.9412	0.8496	51
Normal	1.0000	0.9815	0.9907	54
Squamousc ell	0.8641	0.9889	0.9223	90

	Precision	*Recall*	*F1-Score*	*Support*
Accuracy			0.9016	315
Macro avg	0.9016	0.9237	0.9072	315
Weighted avg	0.9128	0.9016	0.9009	315

CONCLUSION AND FUTURE SCOPE

CT scan images helps in predicting whether a person has lung cancer or not. TL has proven to be effective in the proposed work for the goal of detecting lung cancer. The model's evaluation parameters include precision, recall, and F1-score. The following values for each dataset: Adenocarcinoma-96.96%, large cell-77.42%, squamous cell- 86.41%, and so on. The proposed model has achieved a predicted accuracy of 90.16% using a pre-trained model. It would be intriguing to research the application of TL to additional lung cancer detection tasks, such as nodule detection and categorization. EfficientNet B2 is relatively small and computationally efficient compared to other high- performing models.

The issue of negative Transfer is the main barrier to Transfer Learning. TL is only efficient if the goal and initial matters are comparable enough for the first Learning to be applicable. The number of trainable parameters will change if the first layers are removed, which will cause problems for your dense layers. Deep convolutional layers and densely connected layers can be advantageous for a reduction, but it might take huge time. The whole processing time for the model is long, and overfitting occurs when the new model learns details and noises from training data that have a detrimental influence on its outputs.

REFERENCES

[1] Causey, Jason, et al. "Spatial Pyramid Pooling With 3D Convolution Improves Lung Cancer Detection." IEEE/ACM Transactions on Computational Biology and

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

Bioinformatics (2000).

[2] Sahu, Satya Prakash, Narendra D. Londhe, and Shrish Verma. "Pulmonary nodule detection in CT images using optimal multilevel thresholds and rule-based filtering." IETE Journal of Research 68.1 (2002): 265-282.

[3] Rani, K. Vijila, and S. Joseph Jawhar. "Novel technology for lung tumor detection using nanoimage." IETE Journal of Research 67.5 (2001): 699-713.

[4] Alzubaidi, Mohammad A., Mwaffaq Otoom, and Hamza Jaradat. "Comprehensive and Comparative Global and Local Feature Extraction Framework for Lung Cancer Detection Using CT Scan Images." IEEE Access 9 (2001): 158140-158154

[5] Ozdemir, Onur, Rebecca L. Russell, and Andrew A. Berlin. "A 3D probabilistic deep learning system for detection and diagnosis of lung cancer using low-dose CT scans." IEEE transactions on medical imaging 39.5 (2019): 1419-1429.

[6] Tan, Aaron C., Johan Chan, and Mustafa Khasraw. "The role of immunotherapy in fusion-driven lung cancer." Expert Review of Anticancer Therapy 21.5 (2001): 461-464.

[7] Senthil Kumar, K., K. Venkatalakshmi, and K. Karthikeyan. "Lung cancer detection using image segmentation by means of various evolutionary algorithms. Comput Math Methods Med 2019: 4909846." (2019).

[8] Neal Joshua, E. S., Bhattacharyya, D., Chakkravarthy, M., & Byun,

[9] Y. C. (2021). 3D CNN with visual insights for early detection of lung cancer using gradient-weighted class activation. Journal of Healthcare Engineering, 2001.

[10] Fan, Xiaojie, Xiaoyu Zhang, Zibo Zhang, and Yifang Jiang. "Deep learning-based identification of spinal metastasis in lung cancer using spectral CT images." Scientific Programming 2001 (2001).

[11] Marappan, Shanmugasundaram, Muhammad Danish Mujib, Adnan Ahmed Siddiqui, Abdul Aziz, Samiullah Khan, and Mahesh Singh.

[12] "Lightweight Deep Learning Classification Model for Identifying Low-Resolution CT Images of Lung Cancer." Computational Intelligence and Neuroscience 2002 (2002).

[13] Wang, Y., Nazir, S., & Shafiq, M. (2001). An overview on analyzing deep learning and transfer learning approaches for health monitoring. Computational and Mathematical Methods in Medicine, 2001.

[14] Noronha, Vanita, et al. "Lung cancer with dual EGFR and ALK driver alterations at

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

baseline: a retrospective observational cohort study." Acta Oncologica 61.9 (2002): 1143- 1147.

[15] Vijay, R., et al. "COVIDPRO-NET: a prognostic tool to detect COVID 19 patients from lung X-ray and CT images using transfer learning and Q-deformed entropy." Journal of Experimental & Theoretical Artificial Intelligence (2021): 1-16.

[16] Nageswaran, S., Arunkumar, G., Bisht, A. K., Mewada, S., Kumar, J.

[17] N. V. R., Jawarneh, M., & Asenso, E. (2002). Lung Cancer Classification and Prediction Using Machine Learning and Image Processing. BioMed Research International, 2002.

[18] Jaszcz, A., Połap, D., & Damaševičius, R. (2002). Lung X-Ray Image Segmentation Using Heuristic Red Fox Optimization Algorithm. Scientific Programming, 2002.

[19] Feng, Jianxin, and Jun Jiang. "Deep Learning-Based Chest CT Image Features in Diagnosis of Lung Cancer." Computational and Mathematical Methods in Medicine 2002 (2002).

[20] Skallevold, H. E., Vallenari, E. M., & Sapkota, D. (2001). Salivary biomarkers in lung cancer. Mediators of Inflammation, 2001.